MULTIMEDIA UNIVERSITY

STUDENT ID NO

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 2, 2019/2020

TDS2101 – INTRODUCTION TO DATA SCIENCE

(All sections / Groups)

2 MARCH 2020
9.00 a.m. – 11.00 a.m.
(2 Hours)

INSTRUCTIONS TO STUDENTS

1. This Question paper consists of 7 pages with 4 Questions only, including the cover page.

2. Attempt **ALL** questions. All questions carry equal marks and the distribution of the marks for each question is given.

3. Please write all your answers in the Answer Booklet provided.

Question 1

a) The increasing variety of Big Data sources have enabled the use of less structured types of data. *Describe* ONE difference between **semi-structured data** and **unstructured data** and *provide* an example for each type of data.

(3 marks)

b) One of the six characteristics of Big Data is *veracity*. From the perspective of an online retail/commerce company, describe ONE example of **high veracity** data and ONE example of **low veracity data** that are used by the company.

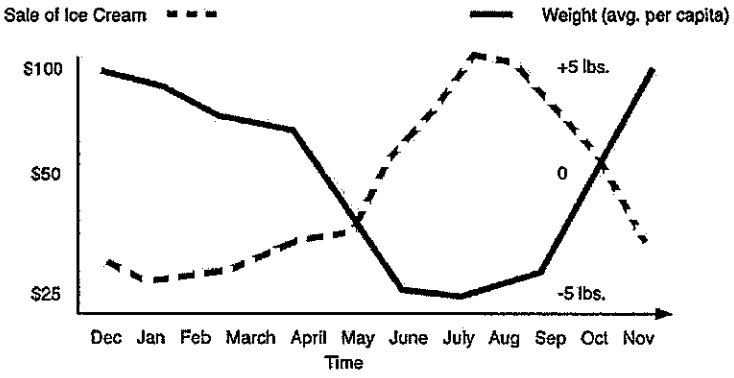
(2 marks)

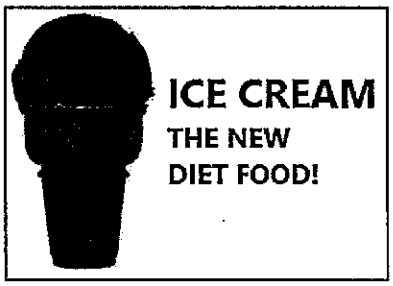
c) A data scientist has obtained climate data containing the following measures (temperature, humidity, precipitation), rice paddy crop yields (metric tonne per hectare produced), and the total rice consumption in the country (metric tonne). Data is available in a 15-year period.

i. If the data scientist wants to study the relationship between the rice consumption and rice paddy yields, what *type of question* (among the six known types) should he/she formulate? (1 mark)

ii. Provide an example of a reasonable **Predictive** question based on the available data. (1 mark)

d) The following analysis of the monthly sales of ice-cream vs. the average per-person weight of the community (*left diagram*) shows that ice-cream is now attributed as the new "diet food". Flyers from ice-cream companies (*right diagram*) have gone out publicizing this newly found conclusion. Provide your own analysis as to whether this claim is reasonable.





(3 marks)

Continued...

Question 2

In recent years, the insurance industry has begun using driving behavior data to determine customer insurance premiums. The following dataset shows a portion of a user's daily driving data collected from a monitoring device that has been fitted in the user's vehicle (a technology called 'mobile telematics'). The device also comes with an algorithm that calculates a Driving Score (ranging from 0-100; 100 for safe driving and 0 otherwise), and retrieves the weather condition from a free web service. In addition, the dataset also compiles optional user feedback through an in-vehicle dashboard app for the user to add remarks.

Id	Date	Total Distance	Travel Duration	Driving Score	Weather	User Remarks
24	8/23/2019	80	1h 24m	96	Fair	
25	8/25/2019	75	0h 57m	58	Overcast	asdfgdfdsdf
26	8/26/2019	12	0h 29m	55	Rain	
27	8/28/2019	38	0h 50m	45	Rain	rainy day
28	30/8/2019	89	1h 13m	90	Fair	
29	31/8/2019		0h 00m	100	Fair	
30	2/9/2019	92	1h 36m	94	Fair	best best ! :)
31	5/9/2019	36	0h 44m	62	Rain	
32	6/9/2019	25	0h 37m		Overcast	

a) Data cleaning is a data wrangling step that involves solving discrepancies resulting from noisy or incomplete data. Identify any errors/inconsistencies that occur in the following attributes and propose a suitable way to resolve it.
   i. Date
   ii. Total Distance

(3 marks)

b) Describe how the following data transformation procedures can be performed on the following attributes:
   i. Transform Travel Duration, to numerical form
   ii. Transform Driving Score, to categorical form

(2 marks)

c) One of the six dimensions of data quality is *accuracy*. Comment on the *accuracy* of TWO attributes collected in this dataset.

(2 marks)

d) There was a device error in calculating the Driving Score in record 32. In order to fill in this missing value, explain the steps in brief as to **how** the Driving Score can be predicted using the available information.

(3 marks)

Continued...

Question 3

a) The CEO of a new telecommunications company U-Celxis is planning to gain an upper hand over its competitors by using data analytics to better understand their customers' behaviour. The following is a sample of the monthly generated summarised customer data (in JSON format) extracted from their database:

```
{       "userId": 1103,
        "location": "Sungai Buloh",
        "annual-income": 86000,
        "number-of-lines": 2,
        "usage": [
                {       "month": "Nov-19",
                        "bill": 35.50,
                        "data-usage": "0.52G"
                        "call-usage": "129m 43s"
                        "sms-usage: 25
                }
                {       "month": "Dec-19",
                        "bill": 42.30,
                        "data-usage": "0.44G"
                        "call-usage": "188m 20s"
                        "sms-usage: 17
                }
        ]
}
```

Describe *how* this data can be used to help the company to profile their users better and provide ONE suggestion on how the company can leverage such insights to benefit its business.

(4 marks)

b) The data analytics team of MyGrozer, an online grocery store, has perform market basket analysis on their product transactions. The following are some parts of the analysis that warrants a closer look. The table on the left shows a list of association rules (unsorted) generated using the apriori algorithm. The table on the right shows the support of individual items in the store.

Itemset	Support	Confidence	Lift
Milk → Cereal	7%	30%	1.3
Milk → Juice	0.5%	2%	0.4
Milk → Bread	0.5%	3%	2.2

Transaction	Support
Milk	12%
Cereal	30%
Juice	7%
Bread	5%

Continued...

i. Which of the following two items can be shown together on the same promotional page of MyGrozer's mobile application?

(1 mark)

ii. *Explain* your answer, by contrasting against another likely itemset which also demonstrated strong association.

(2 marks)

c) The following table shows a *confusion matrix* which contains the results of running model X to predict the onset of diabetes from 1,300 patients who visited hospitals nationwide.

		Predicted	
		Yes	**No**
Actual	**Yes**	524	126
	No	241	409

i. Calculate the *Accuracy* of performing prediction with model X. Round up your answer to 2 decimal places.

(1.5 marks)

ii. Is the *accuracy* performance metric a suitable measure to summarise the performance of this prediction task? Provide a reason to support your answer.

(1.5 marks)

Continued...

Question 4

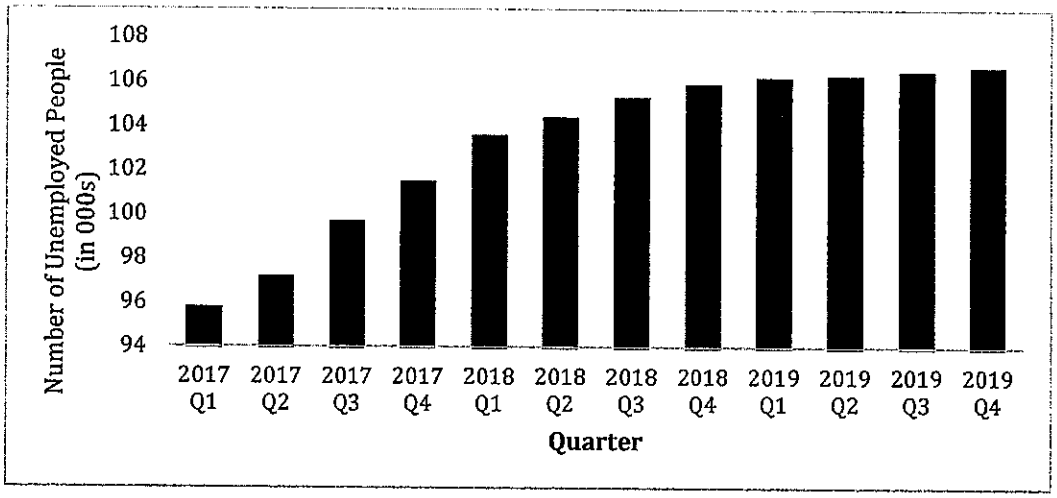
a) Write a line of Python code to perform the following operations on a pandas Dataframe `df` (shown below), consisting of 5 columns of data. Assume that the values of all columns are numeric except for Country. Note: All operations are independent of each other. You do not need to write a full program.

	Country	Year	GDP per capita	Life Expectancy	Education Score
0	Malaysia	2017	11720	74.5	54.9
1	Malaysia	2018	12210	74.5	55.5
2	Malaysia	2015	10524	74.6	52.3
3	Malaysia	2016	11220	74.4	54.2
4	Vietnam	2018	1964	80.9	44.3
5	Vietnam	2017	1853	80.3	41.2
...	...	...	...	...	...
159	South Korea	2015	24870	82.1	72.3

i. Retrieve the Education Score column from the Dataframe.
ii. Group the Dataframe based on Year, aggregation done on other numeric values by taking the mean.
iii. Sort the Dataframe based on Country, then Year, in descending order.

(3 marks)

b) The Department of Statistics of Wakanda has disputed the latest report from the job survey company Panther Inc., which claimed that the number of unemployed people in the country with a population of 50 million, has been "rampantly increasing out of control". In actual fact, the increase has slowed down in the recent years. The following bar chart shows the number of unemployed people in the last three years, which was released together with the report.



Continued...

i. Why does the bar chart visualization above may potentially convey a misleading message?

(2 marks)

ii. Suggest TWO ways to improve this visualization *visually*.

(2 marks)

c) An important aspect in a data scientist's professional code of conduct relates to the **quality of data** and **quality of evidence**. Describe TWO points that data scientists should convey to the client with respect to this aspect of their work.

(3 marks)

End of Paper